Challenges in anaphora resolution in the News Media Genre

P. Nand
School of Computer and Information Sciences
Auckland University of Technology
Auckland, New Zealand
parma.nand@aut.ac.nz

Abstract - *This paper discusses the characteristics of anaphora found in written News Media genre and describes an initial implementation of an algorithm to resolve the anaphors. The paper forms part of wider research aimed at using resolved anaphors for text visualization. The algorithm described is light weight and incremental, in that, it builds vocabulary as it processes documents to be used in future. The input data is also tested on two publicly available anaphora resolution algorithms and the results compared to the algorithm described in this paper. Finally it discusses the challenges in developing light weight algorithms to resolve anaphors in the News Media genre and further research to overcome some of these challenges.*

Keywords: Anaphora, Natural Language Processing, Resolution, antecedent, Ontology, Pleonastic, Noun

1 Introduction

Natural language processing has played a significant role in very diverse fields such as artificial Intelligence, linguistics, philosophy and psychology since about early 1900's. With rapid developments in technology it has become an even hotter topic for researchers with a race towards building computers interacting using natural language. Anaphora resolution plays a key role in natural language processing, helping to build semantics out of written text or spoken discourse.

An anaphor is a word that refers to an entity that has been introduced previously. The most common anaphors in text are pronouns. Consider the sentence:
(S1) John likes apples and he eats them often.

The word "*he*" and "*them*" are pronominal anaphors that refer to the words "John" and "apples" respectively.

Unlike humans, who can resolve anaphors without much difficulty, it is considered a very difficult task for a computer/intelligent system due to inherent knowledge used in the resolution process. Various anaphora resolution models (targeting different types of anaphora) have been developed over the years, however further research needs to be done in order to develop a powerful algorithm that could resolve most, if not all, anaphora. An intelligent system with the capability to do anaphora resolution in any domain would most definitely need to be a knowledge-intensive system. A lot of research has been done and continues to be done where accumulated knowledge is used to resolve anaphors with various degrees of success. However, Lappin and Leass [4] and Mitkov [7] implemented different versions of knowledge-poor algorithms and achieved reasonable results albeit in a limited genre. This research aims to better the performance or knowledge-poor approach on a different genre of printed News Media and apply the results for a practical application.

Anaphora resolution algorithms can be categorised into generally three strategies of attack :

1. Those that apply a set of syntactic rules for resolution [1].
2. Those that apply a set of salience values to candidate antecedents, [3] and [7].
3. Those that use statistical characteristics of candidate antecedents for resolution [2].

Most of the researchers use one or more of the above strategies in combination, to fine tune their algorithms for a particular genre.

2 Comparison of strategies

The approach taken by this research is different to any of the other ones in the reference list in that it is completely based on salience scores yet uses syntactic rules of the language. It is a hybrid which is different to Baldwin [1] and Preiss [8] who also use various forms of hybrids as strategies for the basis of their algorithms. Kennedy and Boguraev [3] use semi-parsed text as the input and allocate positional numbers called offsets for nouns. They don't use information about gender, number and person compatibility in the salience value calculation. Mitkov [7] on the other hand uses number and gender agreement to eliminate unsuitable candidates before applying salience weights using various factors. He used a pre-drawn list of nouns to determine number, person and gender of nouns used in the elimination of the unsuitable candidates.

The strategy taken by this research uses salience weights for the candidates in semi-parsed sentences like Kennedy and Boguraev [3] and Mitkov [7] but instead of eliminating a candidate completely using person, noun and gender criteria it punishes it by being giving it a low salience score hence is still in the running and can be used in case of unavailability of a better suited candidate. The reasoning behind this approach is that we do not use any pre-drawn list of nouns with known person, gender and number values. Instead, this information is extracted from the context of the nouns and rules of the language. Hence, in the cases where the value/s of person, gender

and number can not be determined, the value is left as "*unknown*" which is associated with a low salience score. It is anticipated that in the future other more complex semantics (apart from person, gender and number) about candidates can later be integrated into the algorithm which will aid in resolving anaphors with greater confidence.

3 Model Description

3.1 Pre-processing

A given plain text document is first parsed by a Universal Grammar Engine (UGE) as described by Yeap [11] and the result is output in a clausal form. This result is then parsed and a document object is built which is used by the anaphor resolver module (JavaAR) to resolve the anaphors and populate the antecedent slots in the document object. The populated document object is then used by the visualisation module to output a graphical representation of the document. This research is concerned only with the anaphora resolution part of the whole project.

3.2 Pre-processing of Nouns

The JavaAR module retains a dictionary of the nouns, for future use, from documents being processed. When a document is run for the first time it stores the nouns and any other information derived for that noun from the context of the document. It is anticipated that the retained information will be useful for future processing when it might not be possible to derive the same information from the context of that document. In this way the dictionary list will become richer and richer as more documents are processed, exactly as a child learns a new language. Currently the information derived from the context about the nouns is the gender, number and person. It is anticipated that other more complex information can be derived about nouns from the context of the documents to model humans. A document signature is kept for the documents processed so that this pre-processing of the nouns is not repeated for repeated processing of the same document. The sentences, clauses, nouns and the total number of objects in the document constitute the document signature. The following list of rules gives the current implementation to derive information about nouns.

For each noun
{
If the Noun does not already exist in the noun list
{
Initialise the human, number and gender values of the nouns to unknown.
Nouns ending with "man" and "men" must be human and male. In addition those ending with "men" must be plural and those ending with "man" must be singular.
Nouns ending with "woman" and "women "must be human and female. In addition those ending with "women" must be plural and those ending with "woman" must be singular.
Nouns with modifier "Mr" must be male, human and singular
Nouns with modifier "Mrs", "Miss", "Ms" must be female, human and singular.
Nouns which are names must be singular.
Nouns containing masculine terms like father, brother, son etc must be human, male and singular
Nouns containing feminine terms like mother, sister, daughter etc must be human, female and singular.
Nouns which are not names and end with "ion" must be non human and singular.
Nouns which are not names and end with "s" but no apostrophe must be plural.
}
}

The following table gives the possible value for each of the properties allocated to nouns and pronouns.

Noun Property	Possible Values
Human	T (true),F (false), U (unknown)
Number	S (singular), P (plural), U (unknown)
Gender	M (male), F (female), N (neutral), U (unknown)
Type	Currently not used

3.3 Anaphora Resolution Algorithm Details

JavaAR is totally based on the salience scores or weights hence no noun is completely eliminated from the running even though it might be punished by being given a low or negative salience score. The candidates are chosen from some fixed subset of previous sentences and salience scores are given to each of the candidates based on criteria described later in this paper. The contributing salience scores are then summed up to set the total score. The highest scoring candidate is chosen as the antecedent at the end.

Given below is the description of the algorithm.

Outer loop
For each pronoun
{
For each the noun in the current clause and 3 previous sentences.
{
Get the total Salience Score
Push the noun in the candidates list
}
Choose the object with the maximum salience score
If (there are more than one candidates with the maximum salience score)
Choose the candidate which is closest to the pronoun as the candidate
Else Choose the candidate with the maximum salience score
If (the candidate chosen is for pronouns "*it*", "*its*", or "*itself*" and the candidate is human) Discard this candidate and assume the pronoun is pleonastic.
}//end of outer loop

3.4 Salience score allocation

The actual values of the salience scores are arbitrary, what is crucial as pointed by Lappin and Leass [4] is the relational structure imposed on the various factors by the chosen values. The relative values used for the factors was determined using linguistic patterns in the written news media genre and also studies done by other researchers (for example Kennedy and Boguraev [3] and Lappin and Leass [4].

Human salience score
If (anaphor is it, its or itself)
{
If (candidate's human is true) score = -10
    If (candidate's human is unknown and number is plural) score = -7
        If (candidate is human) score = -5
        If (candidate's human is false and number is singular) score = 5
}
Else
{
If (candidate's human matches with anaphors human) score = 6
If (candidate's human is unknown) score = 1
        If (candidate's human does not match with anaphors human) score = 0
}

Number salience score
If (candidate's number matches with anaphors number) score = 10
If (candidate's human is unknown) score = 1
If (candidate's number does not match with anaphors number) score = -10

Gender salience score
If (candidate's gender matches with anaphors gender) score = 8
If (candidate's gender is unknown and anaphors gender is neutral) score = 5
If (candidate's gender is unknown) score = 3
If (candidate's gender does not match with anaphors gender) score = 0

Syntactic slot salience score
If (candidate's slot is subject) score = 8
If (candidate's slot is object) score = 2
If (candidate's slot is indirect object) score = 1
If (candidate's slot is verb) score = 1
If (candidate is the head object of the slot) score += 2

The total salience score for each noun was calculated to be the sum of 4 salience scores. These scores give the range of total salience scores between -40 and 40. The candidate with the maximum salience score was then chosen to be the antecedent. In the case of a tie in the salience score, the candidate with the shortest distance to the pronoun was chosen. The distance to the pronoun was calculated from the location index on the noun in the list of the nouns in the document object. This list contains nouns in the order in which they appear in the article as it is read from the beginning of the article.

4 Results

The analysis was done with two online articles from a newspaper published in Auckland New Zealand, *The New Zealand Herald*. The articles were named according to the main Name in the article and hence are called *Zaoui* and *Berry* articles.

The articles contained the following categories of anaphora.

Pronominal Anaphora – this is the most common form of anaphor and is fortunately the easiest to resolve. In this category a pronoun co-refers to its antecedent which might either be in the same sentence or in one of the sentences before the pronoun is mentioned.

Definite Noun Phrase anaphor – the anaphor refers to a previously introduced noun phrase

Ontology based anaphor – this is the most difficult category of anaphor where the anaphor refers to some real world knowledge which has not been mentioned previously anywhere in the article.

Pleonastic anaphor – is also known as null anaphor. In this category the pronouns "*it*," *it's"* and "*itself"* refer to nothing in particular. For example the "*it*" in the sentence "*It might rain tonight*"

Reader/Writer anaphor – the anaphor refers to the person consuming the discourse. In the case of an article it might refer to the reader. For example "*you*" in the sentence "*If you pay peanuts you get monkeys*".

The following table summarises the various characteristics of the two articles. The numbers in the brackets are the figures for the correctly resolved anaphors by JavaAR.

Article property	*Zaoui article*	*Berry article*
no. of Sentences	21	37
no of Clauses	62	69
no. of Nouns	162	142
no. of anaphors	32 (22)	35 (12)
no. of pronominal anaphor	26 (21)	24 (9)
no. of definite noun phrase anaphors	1 (0)	8 (5)
no. of ontology based anaphors	1 (0)	0
No. of pleonastic anaphors	2 (1)	3 (2)
No. of reader/writer anaphors	2 (0)	0

It can be seen that JavaAR was able to resolve 22 out of 32 pronominal anaphors for the *Zaoui* article and 12 out of 35 for the *Berry* article. JavaAR was also able to solve 5 definite noun phrase anaphors (which are considered difficult) for the Berry article which were mainly connectives.

The input data was also tested with two other anaphora resolution systems available on the web, namely JavaRap and Mars which are based on algorithms described in articles [7] and [5] respectively. The table below summarises the results of the comparison.

	Zaoui article	*Berry article*
JavaAR	72%	46%
JavaRap	22%	20%
Mars	13%	17%

From the table above it be seen that the initial implementation of JavaAR produces far better resolution rates compared to JavaRap and Mars for the two newspaper articles which were used as the input data for this experiment.

The table below shows some information about the salience scores for the resolved anaphors and the competing candidates.

	Av. Sal. Ante-ceden t	Av. Sal. for cand-idates	Av. No. of cand-idates	Av. Diff in the ante-cedent & candi-date sal.	% resolved
Berry Article	29.74	15.57	6.6	14.17	46
Zaoui Article	29.94	21.38	6.72	8.56	72

Although two articles is not a reasonable sample it can be seen that in the case when the difference in salience values of the competing candidates is smaller the performance was better at 72 % compared to 46%. This also corresponds with the Salience values for the candidates for the Zaoui article being higher which can be attributed to more semantic knowledge present in the Zaoui article leading to better resolution of the anaphors. The amount of information that can be extracted from articles depends on both the genre of the articles as well as individual writing styles of authors.

For the two articles considered for this experiment JavaAR could resolve 3 of the 5 pleonastic anaphors. The simple rule used for resolving pleonastic anaphora was that if the candidate chosen for any of the "it" family of pronouns was a human, then discard the candidate and assume the antecedent to be pleonastic. There were a total of 3 anaphors (ontology based and reader/writer) which had external antecedents relying on contextual knowledge hence can not be reasonably expected to be resolved by a computer system.

Out of the 50 pronominal anaphors between the two articles, JavaAR system incorrectly resolved 20 anaphors. An interesting fact regarding the incorrectly resolved anaphors was that 70% of the incorrectly resolved anaphors were resolved to the correct pronoun, but because the pronoun itself was incorrectly resolved, this error was propagated through to all the subsequent resolutions.

5 Discussion and further work

Currently the input to JavaAR is parsed text from a parser which is part of a separate project. The parser may not be able to parse all the sentences in an article and hence the sentences which are not able to be parsed are ignored. There is a chance that the antecedent for an anaphor might be in the sentence that was not parsed, in which case JavaAR can not be expected to correctly resolve that anaphor. It is envisaged that in future, even after further work to improve the parsing rate of the parser, it may not be possible to parse all the sentences for all the articles. Taking this into consideration JavaAR will be modified to also search for candidates in non-parsed sentences in the test data.

The extraction of semantics from articles can be further developed into more complex derivations to reflect the workings of human mind. Currently JavaAR retains the knowledge about pronouns (eg. gender, human etc.) from articles as it processes them. The value for a property once established is fixed and cannot be changed making it static. Due to numerous exceptions to general linguistic rules it is possible to derive incorrect values for a noun from the context of an article which can be corrected from processing of future articles. This element of self correction can also be incorporated into the JavaAR system. Further, a web interface can also be developed, so that this knowledge could be centrally located which would make JavaAR get "trained" even quicker by being used by more researchers.

It was also found that more than 80% of the antecedents were found to be in the same sentence. Currently JavaAR statically searches the last 4 sentences for possible candidates. Instead, it could be modified to take advantage of the fact that there is a higher chance for finding the antecedent in the same sentence, may be by using salience weights. With the current architecture more salience weights can be added to take advantage of new emerging patterns as and when they become available. The Java AR design is based on salience weights, very similar to neural networks. This enables one to incorporate new facts and patterns as they become available without the need to drastically alter the existing modules. This makes JavaAR extensible and incremental.

6 Conclusion

JavaAR is able to achieve reasonable resolution on a first implementation in the newspaper genre. Its architecture is simple based only on salience weights and can be extended to incorporate new patterns as they become apparent. JavaAR also needs to incorporate more linguistic rules to further derive non-contextual knowledge about nouns. Apart from the contextual knowledge that humans use to resolve antecedents for pronouns, humans also use rules of a language, which can be incorporated into a computer system.

JavaAR also needs to be modified to take advantage of partially parsed articles as this is closer to reality.

7 References

[1] B. Baldwn, CogNIAC: High precision coreference with limited knowledge and linguistic resources, *Proceedings ACL '97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, Madrid, Spain, 1997, pp. 38-45.

[2] N. Ge, J. Hale and E. Charniak, A statistical approach to anaphora resolution, *Proceedings of the Sixth Workshop on Very Large Corpora, COLING-ACL '98*, Montreal, Canada,pp 161-170.

[3] [C. Kennedy, B. Boguraev, Anaphora for everyone: Pronominal anaphora resolution without a paser, *Proceedings 16th International Conference on Computational Linguistics (COLING'96)*, Copenhagen, Denmark, 1996, pp. 113-118.

[4] [S. Lappin, H. Leass, An algorithm for pronominal anaphora resolution*, Computational Linguistics 20(4) 1994*, 535-561.

[5] R. Mitkov, Anaphora Resolution, *Pearson Education Limited, Longman, UK*, 2002.

[6] R. Mitkov, R. Evans, C. Orasan, A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method, *Proceedings CICLing-2002,* Mexico City, Mexico, 2002,168-186

[7] R. Mitkov, Robust pronoun resolution with limited knowledge, *Proceedings COLING'98/ACL'98*, Montreal, Canada, 1998, 869-875.

[8] Preiss Judita, Anaphora Resolution with Memory Based Learning, in *Proceedings of CLUK5*, 2002, pages 1-9.

[9] L. Qiu, M. Y. Kan and T. S. Chua, A public reference implementation of the RAP anaphora resolution algorithm, *Proceedings Fourth International Conference on Language Resources and Evaluation (LREC 2004*, Lisbon, Portugal, 2004, 291-294.

[10] J. R. Tetreault, analysis of Syntax-based pronoun resolution methods, *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL '99)*, Maryland, USA, pp 602-605.

[11] W Yeap, H. Ho, K. Min, P.Reedy, On the Implementation of an Anaphora Resolution System for SmartINFO. Submitted for publication in *Artificial Intelligence*.